Regression

Multi Linear Regression

Assignment 3

BP: Consider only the below columns and prepare a prediction model for predicting Price.

Corolla<Corolla[c("Price","Age\_08\_04","KM","HP","cc","Doors","Gears","Quarterly\_Tax","Weight")]

Now, from the given data we can observe that there is qualitative data after analysing the data we can eliminate the qualitative data as it doesn’t have any effect on the output.  
This can be done by using dummies and dummy packages.

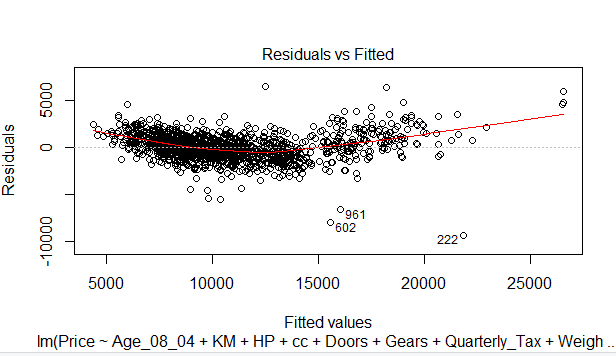
1. Now we have to create a multi linear model for this , by the formula :

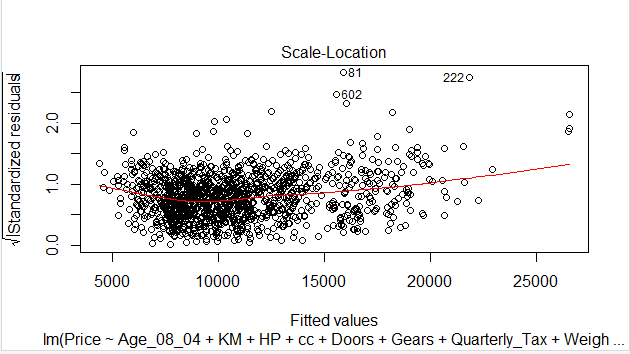
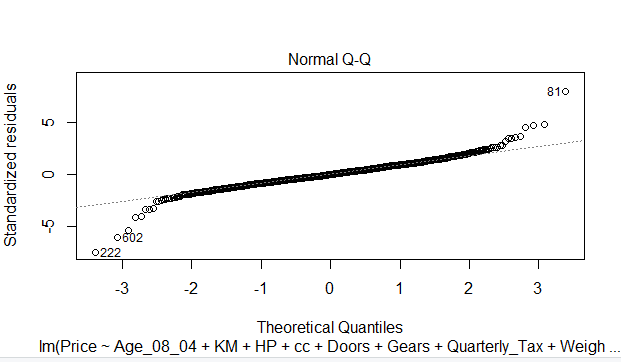
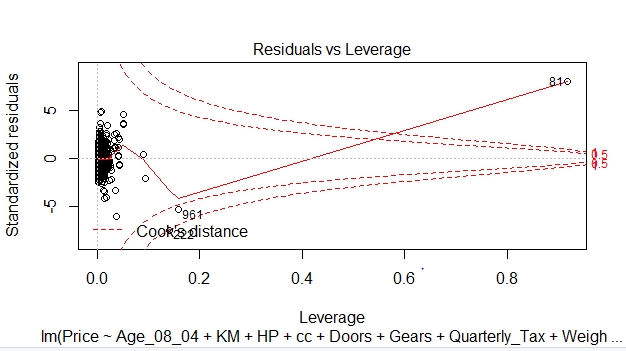
m1 <- lm(price ~ hd + ram + cd + premium + screen+ads+premium + trend,data=cd1)

on using summary ( m1) we get,

Now and value of R2= 0.8638

EDA of the data is as follows:

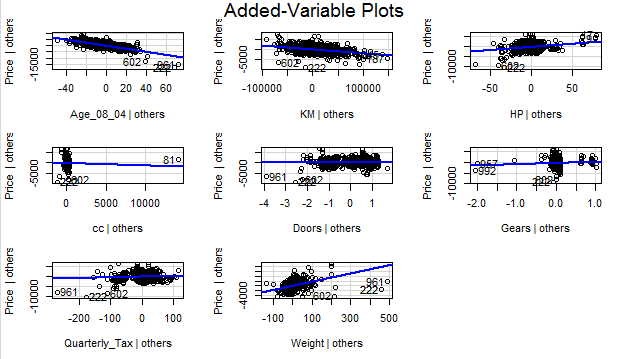


1. Now we have to try and build model , we can use variance inflation models,

If vif > 10 collinearity is good, but from the given values we can see that collinearity is on lower side , we cannot remove all, but can eliminate at least the least one.  
So let’s check which factor is effecting the correlation to bring it down.

Now on plotting added variable plots, we get as follows



From the following plots we can assume that doors , cc has been pulling down the value of correlation . But we cannot judge directly like that , so we use another method of regression called as Akaike information criterion model .  
  
on doing so we get as follows,

Start: AIC=20693.89

Price ~ Age\_08\_04 + KM + HP + cc + Doors + Gears + Quarterly\_Tax +

Weight

Df Sum of Sq RSS AIC

- Doors 1 2943 2571786477 20692

- cc 1 3256511 2575040045 20694

<none> 2571783534 20694

- Quarterly\_Tax 1 16377633 2588161166 20701

- Gears 1 16393629 2588177163 20701

- HP 1 227730786 2799514319 20814

- Weight 1 454465243 3026248777 20926

- KM 1 497917334 3069700867 20946

- Age\_08\_04 1 3898860600 6470644134 22017

Step: AIC=20691.89

Price ~ Age\_08\_04 + KM + HP + cc + Gears + Quarterly\_Tax + Weight

Df Sum of Sq RSS AIC

- cc 1 3254209 2575040686 20692

<none> 2571786477 20692

- Quarterly\_Tax 1 16503849 2588290326 20699

- Gears 1 17093855 2588880332 20699

- HP 1 228761929 2800548406 20812

- Weight 1 484447009 3056233485 20938

- KM 1 498427860 3070214337 20944

- Age\_08\_04 1 3898877516 6470663993 22015

Step: AIC=20691.7

Price ~ Age\_08\_04 + KM + HP + Gears + Quarterly\_Tax + Weight

Df Sum of Sq RSS AIC

<none> 2575040686 20692

- Quarterly\_Tax 1 14976762 2590017448 20698

- Gears 1 17276597 2592317283 20699

- HP 1 225684613 2800725299 20810

- Weight 1 484245502 3059286188 20937

- KM 1 506728527 3081769213 20948

- Age\_08\_04 1 3902107988 6477148674 22014

Call:

lm(formula = Price ~ Age\_08\_04 + KM + HP + Gears + Quarterly\_Tax +

Weight, data = ToyotaCorolla)

Coefficients:

(Intercept) Age\_08\_04 KM HP Gears

-5.478e+03 -1.217e+02 -2.094e-02 3.133e+01 5.990e+02

Quarterly\_Tax Weight

3.737e+00 1.673e+01

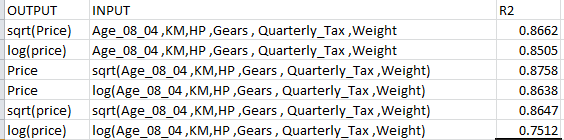
According to AIC regression, it is suggesting to remove cc and doors as we expected .

The value of R2 is = 0.8638.

No change in the value of R2.

To increase the value of R2 we can try some transformations.

The value before and after transformation are:

  
  
So we can conclude that after transformations, square root in the input is working good as we can see increase in R2 value up to 0.8758 which is pity decent.

So we can conclude that from domain knowledge , with increase in age and KM price of car is decreasing and with increase in HP ,Quarterly Tax and weight the price of the car is increasing ( which can be clearly seen in the plots.).